DOCUMENT RESUME

ED 189 192 TM 800 374

AUTHOR Popham, W. James: Lindheim, Elaine
TITLE The Practical Side of Criterion-Referenced Test Development.
INSTITUTION National Council on Measurement in Education, East Lansing, Mich.
PUB DATE 80
NOTE 9p.
AVAILABLE FROM National Council on Measurement in Education, 1230 17th Street, N.W., Washington, DC 20036 ($2.50).
JOURNAL CIT NCME Measurement in Education: v10 n4 p1-8, Spr 1980

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests: *Guidelines: Item Banks: Skill Analysis: *Test Construction: Test Items: Test Validity
IDENTIFIERS *Instructional Objectives Exchange: Test Specifications

ABSTRACT

Attention is drawn to the dynamics of criterion-referenced test (CRT) construction in this report. How CRT's are developed at the Instructional Objectives Exchange is described through a series of three steps. The procedures pertain to the construction of "off the shelf" as well as "customized" tests. Step one, isolating the attribute to be assessed, involves consideration of the skills or attitudinal characteristics proposed for measurement. Each characteristic must satisfy two key criteria: appropriate scope and instructability. Step two, writing the test specification, requires use of a test specification form consisting of five parts. The first four are common to all sets of test specifications: general description, sample item, stimulus attributes, response attitudes. The fifth part, content supplement, is optional. After drafting, the specifications are evaluated for comprehensiveness and precision. Step three, generating the item pool, requires the creation of an item-sampling plan. Variables of concern are the content and subskill dimensions that can differ among items. After identifying the variables, the sample plan is implemented (assigned to different writers), edited, and reviewed. Finally, all CRT items are field tested and judged on the basis of empirical data. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
NCME
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Volume 10, No. 4 Spring, 1980

The Practical Side of Criterion-Referenced Test Development

W. James Popham and Elaine Lindheim

About This Report

W. James Popham

Elaine Lindheim

Much has been written in this past decade about criterion-referenced tests, and much of what has been authored has been heavily influenced by the work of W. James Popham. *Measurement in Education* is pleased, therefore, to present to its readers the following paper, co-authored by Popham and Lindheim.

When first the advocates of criterion-referenced measurement held forth, they tended to do so by contrasting criterion-referenced tests with norm-referenced tests and always presented CRT's as superior to NRT's. There was, unfortunately, not enough light shed on the difference between the narrower-range CRT and the wider-range NRT. The major emphasis appeared to be placed more on the mode of reporting test scores than on the construction of the tests themselves. Popham and Lindheim call our attention to the dynamics of criterion-referenced test construction; they share with us, in a sense, the anatomy of a criterion-referenced test. While the description that follows details how CRT's are developed at the Instructional Objectives Exchange (IOX), it presents as good an analysis of criterion-referenced test construction as this editor has read.

Elaine Lindheim is Director of Test Development at IOX. While she is new to the pages of *ME*, we can look forward to more of her work in the future. W. James Popham has written extensively and is no stranger to the educational measurement community. He is a former president of AERA and is the Director of IOX.

HCR

One would like to assume that most educators are well versed in the steps necessary to produce high quality norm-referenced tests. After all, we've had over half a century's experience in building such tests. Moreover, for several decades we've had access to a host of introductory measurement textbooks replete with recipes as to the birth and nurturing of norm-referenced testing devices.

The availability of a solid, experienced-based technology for creating norm-referenced measures, coupled with effective dissemination of this technical information through conventional textbook channels, has contributed to the substantial number of excellent norm-referenced tests currently at the disposal of educators. Putting it another way, two reasons for the fact that educators can usually find one or more first-rate norm-referenced tests for their purposes are (1) the substantial technical base underlying norm-referenced test construction, and (2) the promulgation of this technology among would-be test

TM 800 374

developers. Yes, although we can always yearn for improvements in quality, when we scrutinize the norm-referenced tests available to today's educators, we come away fairly well satisfied.

But such is not the case with respect to criterion-referenced tests. Recent reviews of the criterion-referenced tests currently available to educators suggest that many of these tests are rather dismal assessment devices.[1] Some are downright despicable. Why, one might wonder, is there such a qualitative disparity between criterion-referenced tests and their norm-referenced counterparts?

One salient reason for the inadequacies in many currently available criterion-referenced tests is that, unlike norm-referenced measurement, we do not possess an experience-honed technology for developing criterion-referenced measures. Furthermore, the modest advances which have been made in the construction of criterion-referenced tests are not well known. For instance, an inspection of currently available educational measurement textbooks will reveal treatments of criterion-referenced test curriculum that could be generously characterized as skimpy.

Whereas norm-referenced test developers have, through the years, shared their insights regarding such prosaic concerns as the care and feeding of multiple-choice alternatives, there are few "show and tell" reports from criterion-referenced test creators.

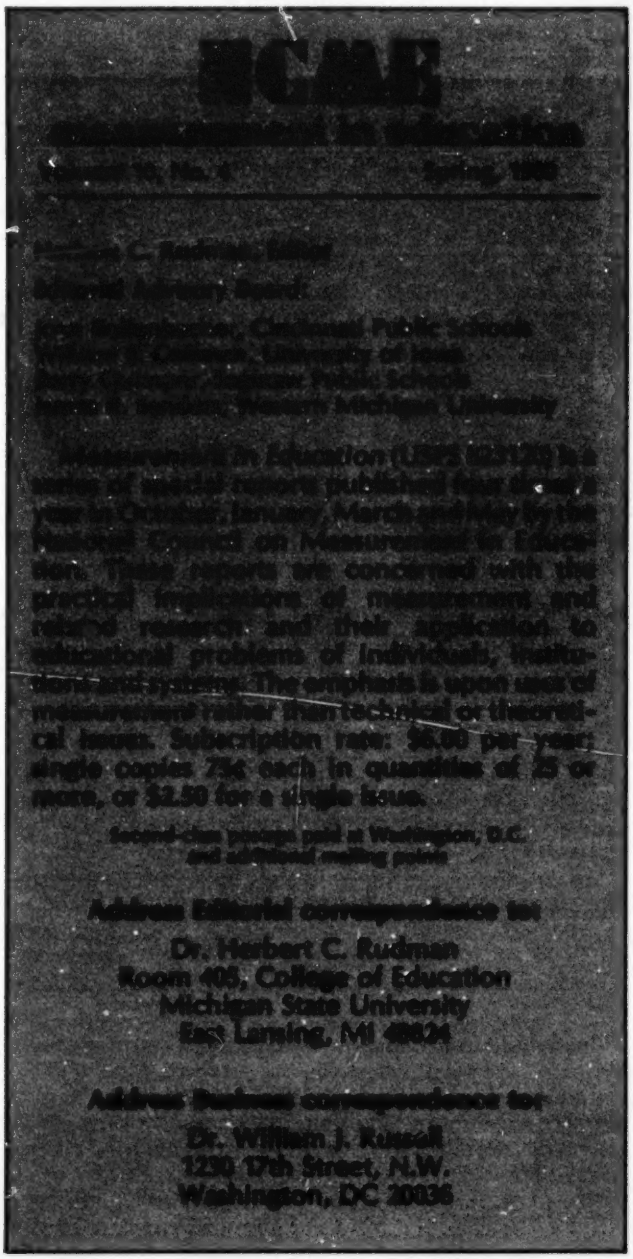
In this report we hope to reverse that trend. Representing an agency that for nearly ten years has been devoted chiefly to the preparation of criterion-referenced measures, we wish to describe some particulars of how we play the criterion-referenced test development game. There is no implication, of course, that this is the one and only way to whip up a winning criterion-referenced test. Departures from the scheme we shall describe are not only permissible, but encouraged. If criterion-referenced test developers try out various approaches, weigh the merits of these alternative procedures, then share their experiences, we shall all profit. In this report we hope to institute a bit of profit-sharing.

More specifically, in the following paragraphs we shall describe the step-by-step procedures employed by the Instructional Objectives Exchange (IOX) in the creation and refinement of criterion-referenced tests. The exposition of these procedures will lean decisively toward the practical. We do not view these pages as an opportunity to delve into the nuances of psychometric theory. But even though practical in its orientation, it must be confessed at the outset that there's more than a little art to the creation of criterion-referenced tests. Ideally, we could bottle our decade's worth of experience, or at least set it out in such replicable form that others would be able to emulate our efforts fairly well. Distressingly, perhaps because of the substantial differences in test content, target examinees, and purpose of every test, each year we find ourselves encountering development problems which are more dissimilar than identical. It will require more than one decade to understand the major idiosyncracies of criterion-referenced test development. Creation of a meaningful technology of criterion-referenced measurement will probably be more difficult and time-consuming than most measurement specialists imagine. You have to live with a centipede for a while before discovering the difficulty of teaching it to samba.

[1] O. Buros, ed. *The Eighth Mental Measurement* (Vols. I & II). Highland Park, N. J.: The Gryphon Press, 1978; or
CSE Criterion-Referenced Test Handbook. CSE Test Evaluation Project. Center for the Study of Evaluation, University of California, Los Angeles, 1979.

Different Flowers in the Garden

Not every flower in the garden is a nasturtium, and not every test called criterion-referenced is a member of the same species. Hambleton, *et al.* have provided a useful



appraisal of competing definitions of criterion-referenced tests.[2] We subscribe to the general conception of a criterion-referenced test proffered there, namely, that it is a test which allows us to describe an examinee's status with respect to a defined behavioral domain. But it is the rigor of the behavioral domain's definition which really allows us to sort out the quality of different criterion-referenced tests.

Some criterion-referenced measures are constructed so cavalierly that the user of the examination is almost at a loss to determine just what it is that the test is measuring. Typical of such exams would be the so-called objective-based tests in which a pool of test items is allegedly "related" to a behavioral objective.

At the other extreme, some criterion-referenced tests are crafted with such care that users can discern with great clarity what behavioral domain the exam sets out to tap. Examples of such highly descriptive examinations would be the domain-referenced tests developed some years ago by Hively and his colleagues.[3] Such tests are accompanied by highly detailed sets of specifications which are employed in the creation of items for the test. In fact, it is these sets of specifications which serve as the descriptive mechanism through which the test's developers communicate with others regarding what it is that the test measures.

It may be helpful to characterize the rigor of a criterion-referenced test's descriptions along a continuum such as that depicted below where we see that at the left we have fairly terse, somewhat loose descriptive information accompanying the test. At the right, we find a far more elaborate and circumscribing set of descriptive information.

Terse and Loose — Detailed and Delimiting

A Continuum of Descriptive Rigor for Criterion-Referenced Tests

Note that it is the detail and the delimitation of the descriptive information that serves as the key distinction among criterion-referenced tests. A particular item on a test, or even a number of such items, fails to indicate the descriptive rigor of the test. Only by consulting the descriptive information that accompanies those items can we discern whether we're dealing with a loosely or tightly defined behavioral domain.

At IOX, we have moved increasingly to the right over the years. More accurately, our descriptive information has become increasingly detailed as we have become more and more experienced in generating criterion-referenced tests. The reason for this movement towards greater descriptiveness is straightforward. We have become increasingly convinced that the major contribution of the criterion-referenced tests is the descriptive lucidity that they provide. Our experiences repeatedly show that to provide a truly illuminating description of examinee behavior requires a greater rather than lesser degree of detail.

Please recognize, therefore, that in the following exposition we shall be describing procedures for growing a particular type of nasturtium. Those preferring geraniums must look in somebody else's flower patch.

Steps in Constructing a Criterion-Referenced Test

We turn now to a step-by-step depiction of how we go about creating a criterion-referenced test at IOX. The procedures to be followed here pertain, incidentally, to both the construction of *off-the-shelf tests* (that is, tests we market directly to purchasers who use such tests "as is") as well as the construction of *customized tests* (that is, tests developed on a contract basis for customers such as school districts or state departments of education).

In the case of the off-the-shelf tests we call all the shots ourselves, attempting to make reasonable decisions about what will best meet educators' needs for criterion-referenced measures. Although we always secure the counsel of practicing educators regarding prospective tests, it is we, not they, who decide whether the tests should measure Skill *A* or Skill *B*, and how many items-per-skill or forms-per-test should be constructed.

In the case of customized test development, on the other hand, it is the customer's call. We try to influence the customer's decisions, of course, but in the final analysis a client decides what behavioral domain should be measured and what sort of test is desired. Nonetheless, even with this external control, we invariably employ a test development scheme akin to the one described below.[4]

Step One: Isolating the Attribute to be Assessed

A test development project begins with a careful consideration of the skills or attitudinal characteristics proposed for measurement. Each one must satisfy two key criteria: namely, *appropriate scope* and *instructability*. Proposed skills or attitudes found deficient are modified until judged adequate before any further development takes place. For purposes of this discussion, we'll be focusing on *skills* to be measured, although one might as readily create tests to assess attitudes or knowledge.

A potentially measurable skill's *scope* is determined by the dimensions of the behavioral domain circumscribed for that skill. In order to get an idea of a proposed skill's scope, we place the skill on a continuum that ranges from "very narrow" to "very broad." A narrow scope skill represents a relatively small behavioral domain. For example, the skill of being able to add pairs of one-digit numbers is a narrow scope skill. Compare it with the skill of being able to add any number of multi-digit integers,

[2] Ronald K. Hambleton, Hariharan Swaminathan, James Algina, and Douglas Bill Coulson, "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Development," *Review of Educational Research*, 48, no. 1 (Winter 1978).

[3] Wells Hively, Graham Maxwell, George Rabehl, Donald Sension, and Stephen Lundin, "Domain-Referenced Curriculum Evaluation: A Technical Handbook and A Case Study From the Minnemast Project," *CSE Monograph Series in Evaluation*, Center for the Study of Evaluation, University of California, Los Angeles, California, 1973.

[4] The scheme is described in greater detail in Popham, W. James, *Criterion-Referenced Measurement*, Prentice-Hall, Englewood Cliffs, N. J., 1978.

fractions, decimals, percents or mixed numbers. This second is far broader in its scope.

Exactly how broad or narrow a skill's scope should be in order for it to be deemed appropriate is almost entirely dependent upon the purposes of the instrument being designed. The only exception to this rule is that a skill that is too broad in its scope will never prove serviceable as a focus for criterion-referenced measurement. Such a skill represents an attempt to coalesce two or more separate, non-combinable skills into a single skill. "Reading comprehension," for example, would be a skill that is too broad in its scope, as there are several essentially discrete intellectual skills represented by that one descriptor.

At IOX we find ourselves increasingly involved with the creation of end-of-instruction competency certification examinations. Fairly broad-scope objectives are appropriate here, with each objective subsuming several enroute, enabling skills. If, on the other hand, we were creating an examination designed to yield detailed diagnostic information about an examinee's specific weaknesses in a subject field, fairly narrow-scoped objectives would be more suitable.

Whereas the criterion of *scope* is a relative one, the criterion of *instructability* is not. In applying this standard, we are not sorting skills into groups of more or less instructable goals. Instead, we are determining whether a skill represents an examinee's behavior that can be influenced by instruction. In effect, we are operationalizing our agency commitment to creating measures that are sensitive to instruction. We want to fashion measures that will serve as agents of instructional improvement. We attempt to do so by creating attainable targets and then clarifying those targets for teachers so that they can be more effective in helping students achieve those goals.

In some instances, a skill itself will clearly be deficient when examined for *instructability*. It may represent a skill that is primarily a function of native intelligence or an attribute that is primarily a reflection of individual temperament. It is not that such skills and behaviors cannot be measured by criterion-referenced tests; it is just that they do not belong in a measurement system oriented around an instructional program.

After a skill has passed the preliminary screens of *appropriate scope* and *instructability*, it is ready to have its attributes specified. The end-product of this step in test development will be a several-page document that defines, as precisely as possible, how the skill will be measured. These test specifications will be used by item writers to generate the items that will measure the skill.

Step Two: Writing the Test Specifications

The test specifications form used at IOX consists of five parts. The first four parts are common to all sets of test specifications and the fifth part (a Content Supplement) is optional, depending upon the attribute being described. Each part follows a certain format and contains a specific type of information. (A sample set of test specifications is presented at the end of this section, to illustrate the points discussed below.)

The first part of a set of test specifications is its *General Description*. This is an overview, or abstract, of everything that follows. The nature of the test question and answer options, as well as the key dimensions that characterize a correct response to the test item, are described. The information here is intended primarily for readers of the specifications who want to get a general idea of the attribute being assessed and the means by which it will be measured. The information is not specific enough to generate test items, but it is sufficient to communicate in an introductory fashion with potential users of the test.

A *Sample Item* follows the general description. This test item exemplifies the type of item that will be generated by the test specifications. It is identical to all the other items that will be written later, and thus provides another preview of the specifications to follow. Item writers and test users, the two principal audiences for test specification, may consult the sample item to gain more insight into the nature of the attribute being measured. Neither audience, however, receives sufficient information from the item alone. Item writers still have not been provided with the rules for generating other items, and test users still have not received guidance for interpreting examinee performance relative to the skill.

The answer choice analyses that accompany the sample item are derived from the criteria used to distinguish correct and incorrect answer options. These criteria are defined in detail in a later portion of the specifications.

The third and fourth parts of the specifications, the *Stimulus Attributes* and *Response Attitudes* components, lay out the rules for test item generation. In developing test specifications, there is a constant tension between the desire to tie down every possible detail associated with the test item form and the reluctance to overwhelm item writers and test users with insurmountable amounts of information. The question of how much detail is enough almost always has to be resolved judgmentally, by giving a set of specifications to individuals who have not been associated with the development of the document, then asking them to write several test items using those specifications. Both the items they write and the comments they make about problems in understanding the specifications will indicate whether the level of detail is appropriate.

Some sets of test specifications contain a *Content Supplement*. A supplement is employed to enumerate specific content eligible for testing; the stimulus and response attributes portions of the specifications are reserved for general rules only. Supplements may include such information as a list of all the words that can be used on a spelling test, a compilation of terminology that may be included in a set of mathematics word problems, or an itemization of the diseases, diagnoses, and patient characteristics eligible for testing in the measurement of a medical skill.

Almost every set of test specifications stipulates that some sort of readability procedure be applied in generating test items. This is the case even if the attribute being assessed is not a reading skill. Readability controls may be used to constrain the overall difficulty of reading passages by means of a standard readability formula that yields a grade equivalent or approximate difficulty range for a selection. Such formulas, however, are not applicable to single sentence questions or answer choices. Here it is often desirable to limit the number of words and the grammatical complexity of the test materials, as well as the level of difficulty of the individual words themselves.

After a set of test specifications has been drafted (a process that can take anywhere from several hours to a month or more, depending upon the complexity of the attribute being described) the specifications are subjected to intensive internal agency review. The specifications are

evaluated for their comprehensiveness, clarity, and precision. If they are being developed under contract, they are also reviewed by the individuals commissioning them. Extensive revisions and substantial improvements often occur as a result of both these review operations. When the specifications are approved, the next stage of test development can begin.

Illustrating Test Specifications
IOX Basic Skill Tests
Secondary Level
Reading
Determining Main Ideas

General Description

The student will be presented with a factual selection such as a newspaper or magazine article or a passage from a consumer guide or general-interest book. After reading that selection, the student will determine which one of four choices contains the best statement of the central point of selection. This statement will be entirely accurate as well as the most comprehensive of the choices given.

Sample Item

Directions: Read the selections in the boxes below.
Select the best main idea statement for each selection.

The Cold Facts

Had you lived in ancient Rome you might have relieved the symptoms of a common cold by sipping a broth made from soaking an onion in warm water. In Colonial America you might have relied on an herbal concoction made from sage, buckthorn, golden-seal, or bloodroot plants. In Grandma's time, lemon and honey was a favorite cold remedy, or in extreme cases, a hot toddy laced with rum.

Today, if you don't have an old reliable remedy to fall back on, you might take one of the thousands of drug preparations available without prescription. Some contain ingredients much like the folk medicines of the past; others are made with complex chemical creations. Old or new, simple or complex, many of these products will relieve some cold symptoms, such as a stopped up nose or a hacking cough. But not a single one of them will prevent, cure, or even shorten the course of the common cold.

1. Which one of the following is the best statement of the main idea of the article you just read?
   a. Old-fashioned herbal remedies are more effective than modern medicines. (lacks accuracy — unsupported)
   b. Throughout history, many remedies but no real cures have been developed for the common cold. (correct)
   c. Some of today's cold preparations contain ingredients much like those found in folk remedies of the past. (lacks appropriate scope — too narrow)
   d. Simple as well as complex drug preparations and folk medicines have been created to treat a wide variety of health problems. (lacks appropriate scope — too broad; lacks accuracy — unsupported)

Stimulus Attributes

1. A test item will consist of a reading selection followed by the question "Which one of the following is the best statement of the main idea of the (article/selection) you just read?" A reading selection will be titled, will be at least one paragraph long, and will contain from 125-250 words.
2. Not more than 1,000 words of reading material will be tested in a set of five items. At least two of the five items in any set of five items will contain reading selections that are more than one paragraph long.
3. Reading selections eligible for testing will be adaptations of passages from factual texts such as newspapers, magazines, general-interest books, and consumer guides. Care will be taken to pick selections of particular interest to young adults and to avoid selections which may appear dated in the near future.
4. A reading selection will communicate one central idea. This idea may or may not be stated explicitly in the selection.
5. If necessary, the following modifications may be made to a selection used for testing:
   a. A title may be added, if the selection does not have one, or if the selection represents a section of a longer piece whose title would not be applicable to the excerpt. If a title is added, it will be composed of a brief, interest-getting and/or summarizing group of words.
   b. A selection may be shortened, but only if the segment which is to be used for testing makes sense and stands as a complete unit of thought without the parts which have been omitted.
6. Reading selections used for testing will be between a seventh grade and a ninth grade reading level, as judged by the Fry readability formula.

Response Attributes

1. A set of four single-sentence response alternatives will follow each reading selection. One of these alternatives will be the correct main idea statement for the selection. The other three will be incorrect main idea statements. All four statements will plausibly relate to the content of the reading selection, either by reiterating or paraphrasing portions of that selection, or by building upon a word or idea contained in the selection.
2. The three incorrect response alternatives will each lack one or both of the two characteristics needed in a correct main idea statement: *accuracy* and *appropriate scope*.
3. A distractor *lacks accuracy* if it:
   a. Makes a statement contradicted by information in the text
   b. Makes a statement unsupported by information in the text (Such a statement would be capable of verification or contradiction if the appropriate information were attainable.)
   c. Makes a statement incapable of verification or contradiction; that is, a statement of opinion (Such statements include value judgments on the importance or worth of anything mentioned in the text.)
4. A distractor *lacks appropriate scope* if it:
   a. Makes a statement that is too narrow in its scope (That is, the statement does not account for all of the important details contained in the text.)
   b. Makes a statement that is too broad in its scope (That is, the statement is more general than it needs to be in order to account for all of the important details contained in the text.)
5. The three distractors for a selection will include at least one statement that lacks accuracy and one statement that lacks appropriate scope. On a given test, between 10 and 20 percent of the distractors will be sentences taken directly from the text.
6. The correct response for a selection will be a main idea statement that is both *accurate* and of the *appropriate scope*. All the information it contains will be verifiable in the selection upon which it is based. It will encompass all of the most important details contained in that selection.
7. The important points to be included in a main idea statement are those details which are emphasized in a selection by structural, semantical, and rhetorical means such as place-

ment in a position of emphasis, repetition, synonymous rephrasing, and elaboration. Whether any given main idea statement contains all of the important points that it should is always debatable rather than indisputable. The nature of the question asked on this test (i.e., select the *best* main idea statement from among those given) attempts to account for this quality of relative rather than absolute correctness.

8. If a sentence in the text itself qualifies as the best main idea statement which can be formulated about a selection, that sentence may be reiterated as a response option. No more than 20 percent of the items on a given test will have as their correct answer a main idea statement which is a direct restatement of a sentence in the text.

Step Three: Generating the Item Pool

Careful planning precedes the generation of test items, in order to ensure that the needs of the development project will be met after the item pool has been created. The types of measures being created, as well as the resources available to create those measures, work together to determine the number of items per set of specifications that will be written. The ideal development situation allows for several field-testings of items before the final measures must be assembled. Such a scheme allows for the revision of items based upon performance data. Time constraints often make such a plan impossible, necessitating the writing of many surplus items before any tryouts, so that there will be enough acceptable items to use intact, after the performance data are analyzed.

In theory, a set of test specifications will generate an item pool as large as the domain of content it describes. Again in theory, that domain can then be sampled in order to create one or more forms of an examination. This theoretical approach of creating an entire item universe before selecting the items for a specific item pool, however, isn't at all practical. In practice, it is necessary to create an item-sampling plan ahead of the fact to determine in advance which items to write. It is first necessary to identify all the variables of concern. These will be the content and subskill dimensions that can differ among items. In the item pool for a mathematics skill, for instance, the type of number used in a problem, the number of digits, and the arithmetic procedures required to solve the problem might all be variables of interest. After all such variables are identified, the sampling plan is implemented in the form of specific assignments to writers. One writer may be able to complete an entire item pool if it is relatively small. Item pools of many items are usually divided among writers.

After all of the items for a given set of specifications have been written, they must pass through a series of editing procedures. First the items are rigorously scrutinized to discern their congruence with the test specifications. Such a congruency review also verifies that the reading level of the items is appropriate according to whatever readability constraints have been set forth in the specifications. Next the items are screened for their quality. It is possible to write items that, even though congruent with test specifications, are stylistically dismal. Such losing items are characterized by pedestrian writing, uninteresting content, and distractors that are sometimes quite dumb. Operating within the constraints of a set of test specifications, it is possible to create a range of items that all test the skill being measured, with some of those items being far more stellar than others. Writers of criterion-referenced tests need not be automatons. Items should be sought that simultaneously satisfy specification constraints, yet sparkle.

One final review process constrains all of the test items. They are all examined for any potentially biasing elements. Such elements are defined as anything in an item that could potentially advantage or disadvantage any subgroup of examinees within the population to be tested. The content employed in an item is scrutinized here, to make sure that it is dependent upon language and experiences held in common by all examinees. Individuals with particular sensitivities to all of the subgroups within the testing population of interest are gathered to screen the items. If any single reviewer expresses a concern about an item, it is modified until it is acceptable.

Field-Testing the Items

One might think that with so much in-house scrutinizing of items (that is, reviewing them for their congruence with specifications, absence of bias, readability level, etc.) there would be little need for trying the items out with living and breathing examinees. After all, how could test items that have been reviewed, refined, and re-reviewed *ad nauseum* turn out to be duds? Easily.

In fairness, it is less easy for an oft-reviewed item to turn out badly than one that's never been held up to inspection. Nonetheless, even seriously scrutinized items can sometimes go sour. In order to detect defective items, all criterion-referenced test items need to be field-tested and judged on the basis of empirical data. Berk, writing in a previous issue of *Measurement in Education*, has set forth a series of data-based indices of item quality which can be used to judge the virtues of criterion-referenced test items.[5] All of the procedures that Berk describes hinge fundamentally on examinee performances under uninstructed and instructed conditions. In some instances, a single group of examinees' pre-instruction and post-instruction performance on an item are contrasted. In other cases, the item performances of two groups, one instructed and the other uninstructed, are compared. Depending on the availability of examinees, and the time requirements involved (since it takes time, in the case of the one group pre- and post-instruction procedure, for instruction to transpire), these indices of item quality can prove most useful in judging the merits of criterion-referenced test items.

Developers of criterion-referenced tests strive to create essentially homogeneous test items by developing a pool of items which is congruent with a set of specifications for a particular behavioral domain. In the early days of criterion-referenced measurement, it was thought that we could create test specifications so well that the items yielded from those specifications would mirror-image each other. Experience has shown us, however, that even a well constructed bundle of criterion-referenced test items, derived from a well constructed set of test specifications, will fail to display *identical* item characteristics. For instance, even in a fairly small scope behavioral domain, such as one's ability to multiply pairs of double-digit numbers, test items will display substantially different *p* values. Nonetheless, the correlation between exami-

[5] Ronald A. Berk, "A Consumer's Guide to Criterion-Referenced Test Item Statistics," *Measurement in Education*, 9, no. 1 (Winter 1978).

nees' performance on an individual item and on the sum of remaining items in the pool is illuminating. If we find, for example, that the bulk of correlations range between .40 and .70, but that the correlations for a few items fall below .20, then those less strongly correlated items warrant further scrutiny.

At IOX we also employ *p* values to get a rough fix on item quality. As before, if we survey a pool of items designed to assess the same attribute (e.g., skill or attitude) and discover several items with aberrant *p* values, then we look at those items more carefully. It should be stressed that, in most cases, items with atypical *p* values are subsequently judged to be perfectly fine, hence remain in the item pool. It usually turns out that our test specifications are unable to sufficiently constrain item characteristics so that only equidifficult items are produced. In a few cases, of course, the items are indeed departures from what was anticipated in the specifications and therefore must be modified or discarded. Occasionally, we discover that the specifications themselves contain the cause for the *p* value atypicality, thus dictating a modification in those specifications, and a subsequent revision of all items thus affected.

In recent years we have also been subjecting almost all of our items to Rasch analyses in order to develop test forms more readily equitable on the basis of Rasch calibrations. As a consequence, items which fail to satisfy Rasch goodness of fit requirements are also excised or modified (then field-tested again).

In review, empirical field-testing of criterion-referenced test items is a useful adjunct to the dominantly judgmental review operations characterizing initial item development. Ideally, development agencies would want to carry out full blown field-tests with items under development, then subject the resulting performance data to a galaxy of analyses. In practice, of course, field-testing and data analyses cost money, hence one's ardor for empirically based sprucing-up of items must always be governed by available financial and personnel resources. At IOX we sometimes, always with substantial regret, simply have to "make do" with modest, far from exhaustive, field-tests.

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available for $2.50 each per single issue, and for 75¢ each in quantities of 25 or more for a single issue.

Vol. 1,	No. 1	*Helping Teachers Use Tests* by Robert L. Thorndike
	No. 2	*Interpreting Achievement Profiles — Uses and Warnings* by Eric F. Gardner
	No. 3	*Mastery Learning and Mastery Testing* by Samuel T. Mayo
	No. 4	*On Reporting Test Results to Community Groups* by Alden W. Badal & Edwin P. Larsen
Vol. 2,	No. 1	*National Assessment Says* by Frank B. Womer
	No. 2	*The PLAN System for Individualizing Education* by John C. Flanagan
	No. 3	*Measurement Aspects of Performance Contracting* by Richard E. Schutz
	No. 4	*The History of Grading Practices* by Louise Witmer Cureton
Vol. 3,	No. 1	*Using Your Achievement Test Score Reports* by Edwin Gary Joselyn & Jack C. Merwin
	No. 2	*An Item Analysis Service for Teachers* by Willard G. Warrington
	No. 3	*On the Reliability of Ratings of Essay Examinations* by William E. Coffman
	No. 4	*Criterion-Referenced Testing in the Classroom* by Peter W. Airasian and George F. Madaus
Vol. 4,	No. 1	*Goals and Objectives in Planning and Evaluation: A Second Generation* by Victor W. Doherty and Walter E. Hathaway
	No. 2	*Career Maturity* by John O. Crites
	No. 3	*Assessing Educational Achievement in the Affective Domain* by Ralph W. Tyler
	No. 4	*The National Test-Equating Study in Reading (The Anchor Test Study)* by Richard M. Jaeger
Vol. 5,	No. 1	*The Tangled Web* by Fred F. Harcleroad
	No. 2	*A Moratorium? What Kind?* by William E. Coffman
	No. 3	*Evaluators, Educators, and the Publics: A Detente?* by William A. Mehrens
	No. 4	*Shall We Get Rid of Grades?* by Robert L. Ebel
Vol. 6,	No. 1	*On Evaluating a Project: Some Practical Suggestions* by John W. Wick
	No. 2	*Measuring Reading Achievement: A Case for Criterion — Referenced Testing and Accountability* by S. Jay Samuels and Glenace E. Edwall
	No. 3	*Security in a Citywide Testing Program* by Anthony J. Polemeni
	No. 4	*Dear Mama: Why don't they love me anymore?* by Thomas J. Fitzgibbon
Vol. 7,	No. 1	*The Discrepancy Evaluation Model (Part I)* by Andres Steinmetz
	No. 2	*The Discrepancy Evaluation Model (Part II)* by Andres Steinmetz
	No. 3	*What Graduate and Professional School Students Think About Admission Tests* by Leonard L. Baird
	No. 4	*The Paradox of Education Testing* by Robert L. Ebel
Vol. 8,	No. 1	*FACE UP: A Framework for Assessing Career Educators Utilization Programs* by Bruce W. Tuckman
	No. 2	*Grade Equivalent Scores* by Gary Echternacht and *If Not Grade Equivalent Scores — Then What?* by Jeanne M. Plas
	No. 3	*Criticisms of Testing: How Mean is the Median?* by Henry S. Dyer
	No. 4	*The Superintendent and Testing: Implications for the Curriculum* by Herbert C. Rudman
Vol. 9,	No. 1	*A Consumer's Guide to Criterion — Referenced Test Item Statistics* by Ronald C. Berk
	No. 2	*Perspective on Intelligence Testing*, by Roger T. Lennon
	No. 3	*Bias*, by Eric F. Gardner
	No. 4	*The Use of Tests in Admissions to Higher Education* by Mary Fruen
Vol. 10,	No. 1	*Some Comments on Criterior-Referenced and Norm-Referenced Achievement Tests* by William A. Mehrens and Robert L. Ebel
	No. 2	*Does the Rasch Model Really Work? A Discussion for Practitioners* by R. Robert Rentz and Charlotte C. Rentz
	No. 3	*A Perspective On The Uses and Misuses of Standardized Achievement Tests* by Peter W. Airasian

Providing Estimates of Test Quality

Psychometricians have spent decades in creating ways that test developers should describe the worth of their wares. Most of these psychometrically annointed procedures center on a test's reliability and validity. Developers of criterion-referenced measures are not absolved of a responsibility to round up evidence regarding the worth of their assessment devices merely because they're dealing with less traditional measuring approaches. And, although the conventional standards of test quality endorsed by the profession for norm-referenced tests may need a bit of modification for certain criterion-referenced measures, by and large these standards are perfectly appropriate for criterion-referenced tests.

We either create tests under contract or develop tests for off-the-shelf sales. With customizing tests, the particular purposes of the customer determine the sorts of reliability and validity analyses to be conducted. For instance, if we are developing competency tests for a city or state school system which plans to use several forms of the test in subsequent years, then our emphasis is on gathering reliability estimates of the equivalency sort. If, on the other hand, a school district wants to administer one form of a test at different times (as a make-up opportunity for absentees) then reliability evidence of the stability sort is more germane. Similarly, validity questions of relevance to the customer's situation are the chief factors influencing our gathering of validity data.

For off-the-shelf tests, however, the development agency cannot predict precisely how its tests will be used. As a consequence, given available financial resources (which invariably fall short of what we wish) we try to secure a variety of reliability and validity data, then make them available to potential users. This is an instance when more is not less, since some users might wish to focus on a test's equivalence reliability, others on its stability, reliability, and so on.

A Costly Enterprise

We have described, in admittedly brief fashion, the evolution of a criterion-referenced test. The attentive reader will have noted similarities to the lives and loves of norm-referenced measures. The extremely attentive reader will have noted one substantial difference, however, between the two test development odysseys. In the case of criterion-referenced test development there is enormously increased attention to the descriptive rigor of the test specifications, congruence of items with specifications, etc. This increased preoccupation with the semantic attributes of the test items and their descriptive materials is intellectually exhausting work, demanding test development personnel with truly superior cerebral skills. To acquire such a brainy group of staff members, then send them through a relentless quest for precisely defined and otherwise defensible test items, costs far more money than most educators realize.

Too many educators, recalling their own days in the classroom when they whipped up a true-false test in milliseconds, think that *all* test development is both easy and inexpensive. Too many educators, accustomed to using nationally distributed tests which are relatively inexpensive because of their widespread distribution, assume that *all* commercially published tests will be low in cost.

But, to the distress of local school personnel who wish to develop criterion-referenced tests, as well as educators who want to buy inexpensive criterion-referenced tests, the fact is that criterion-referenced tests of reasonably high quality, because of their distinctive emphasis on verbal descriptiveness, will cost more than most people think. Criterion-referenced tests can be produced inexpensively, but they will be of little value either for instructional design or evaluation. Such inexpensively and poorly developed criterion-referenced tests will not yield the very dividends for which those tests were originally conceived.

NCME NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

USPS 823120

1230 17th Street, N.W.
Washington, DC 20036

Second class postage paid
at Washington, D.C.
and
additional mailing points.